# Fundamental of Data Science

Fedlu Nurhussien

(Ass. Proff in computer Science , ASTU,UOG,EIC)

email:abuesmael2005@gmail.com

# Course Outline

- Over view of Data Science
  - Data Revolution
  - Causes of the Data Revolution
  - Demand for Data Science and data related professions
  - Definition of data science
  - The Data Science Discipline Venn Diagram
  - The Identity of Data Science Discipline
  - Skills of data scientists
- DS and its relation with other disciplines
  - AI-Vs Data science
  - Statistics
  - Machine Learning
  - Data Mining
  - Database System

# Course Outline

- Sub Domains (Knowledge areas) of Data Science
    - Data Analytics
    - Data Engineering
    - Data Management and governance
- Data Science Technologies and Tools
- Research Areas in DS
- Python for data analytics
    - Data type and operator
    - Popular packages
    - Data preprocessing
- Introduction to machine learning
- Supervised vs unsupervised learning

# Course Evaluation

- Project 50%
- presentation 10%
- Exam 40%

# Chapter one

# Over view of Data Science

# Data Revolution

- Data is created constantly, and at an ever-increasing rate
- Massive amounts of data about many aspects of our lives
  - Shopping, communication, listening to music, searching for information, expressing our opinions
  - The finance, the medical industry, government, education, retail….
  - Websites track every user's on every click.
  - Smartphone are building up a record of our location
  - Smart cars collect driving habits, smart homes collect living habits, and smart marketers collect purchasing habits.
  - Cross-referenced encyclopedia; domain-specific databases about d/t things

# Data Revolution

- There is a growing influence of data in most sectors and most industries.
- Culturally saturated feedback loop where our behavior changes the product and the product changes our behavior
- **Technology makes this possible**:
  - infrastructure for large-scale data processing,
  - increased memory, and bandwidth, as well as a cultural acceptance of technology

# Big Data - a tsunami that is hitting us

- We are witnessing a tsunami of data:
  - Huge volumes
  - Data of different types and formats
  - Impacting the business at new and ever increasing speeds
- The challenges:
  - Capturing/collecting data
  - Managing
  - Storing - safeguarding and securing
  - Processing

"Big Data refers to non-conventional strategies and innovative technologies used by businesses and organizations to capture, manage, process, and make sense of a large volume of data"

# Data has an intrinsic property…it grows and grows

**90%**

of the world's data was created in the last two years

**80%**

of the world's data today is unstructured

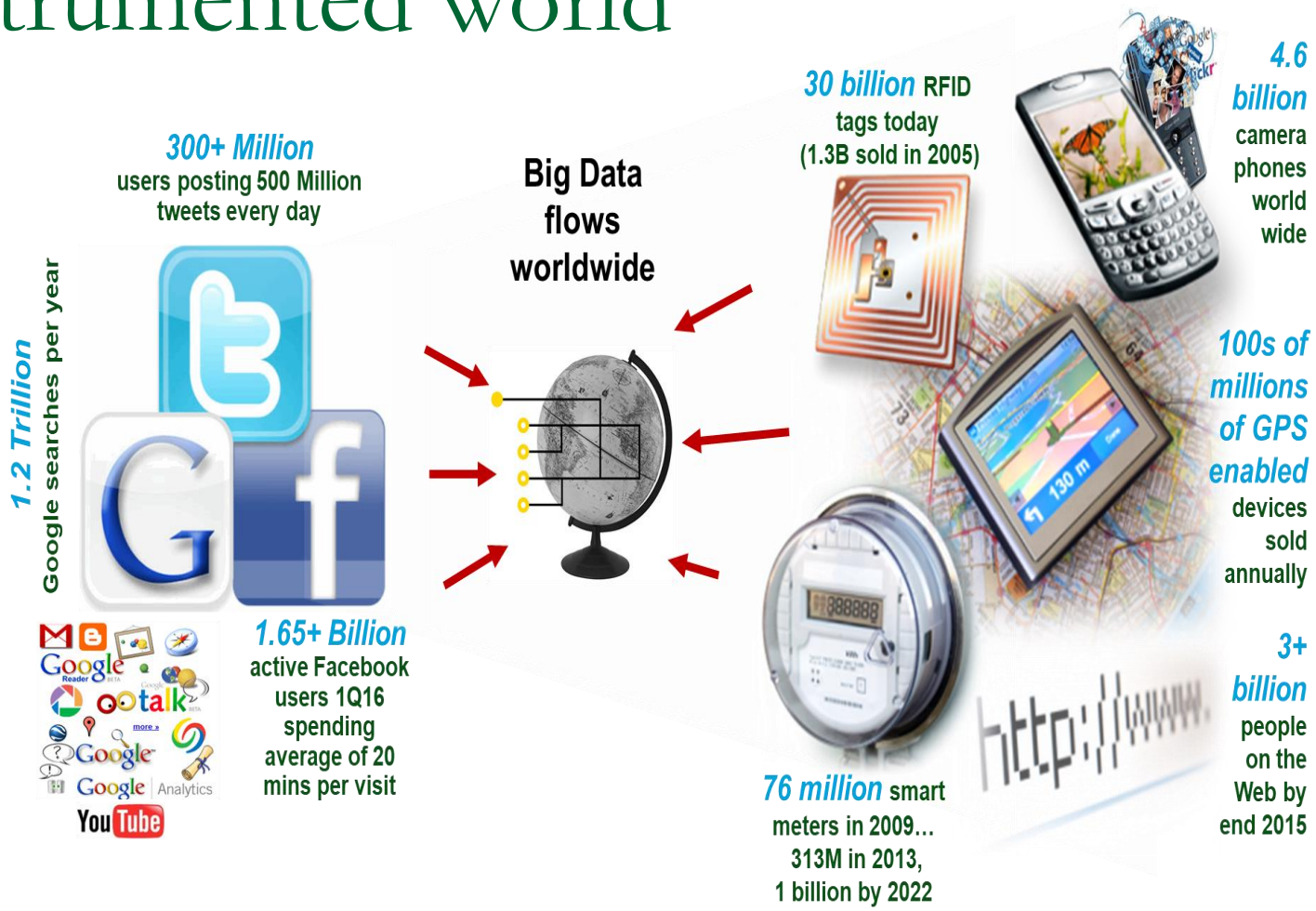
**20%**

of available data can be processed by traditional systems

**1 in 2**

business leaders don't have access to data they need

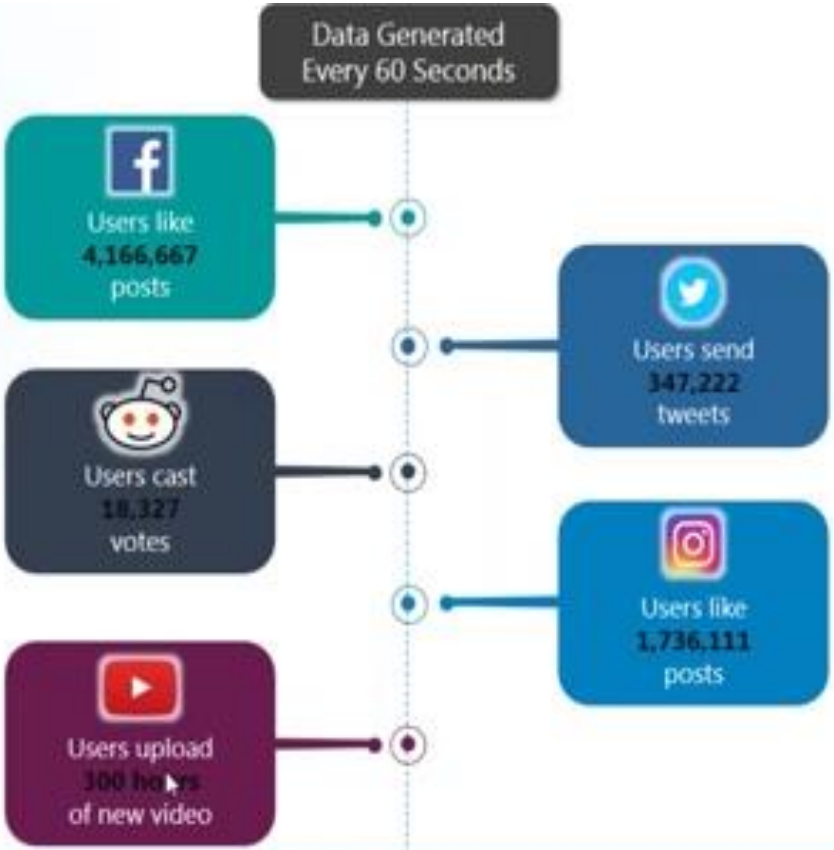
# Growing interconnected & instrumented world




3 billion Facebook users and 8 billon mobile devices

# Data Revolution

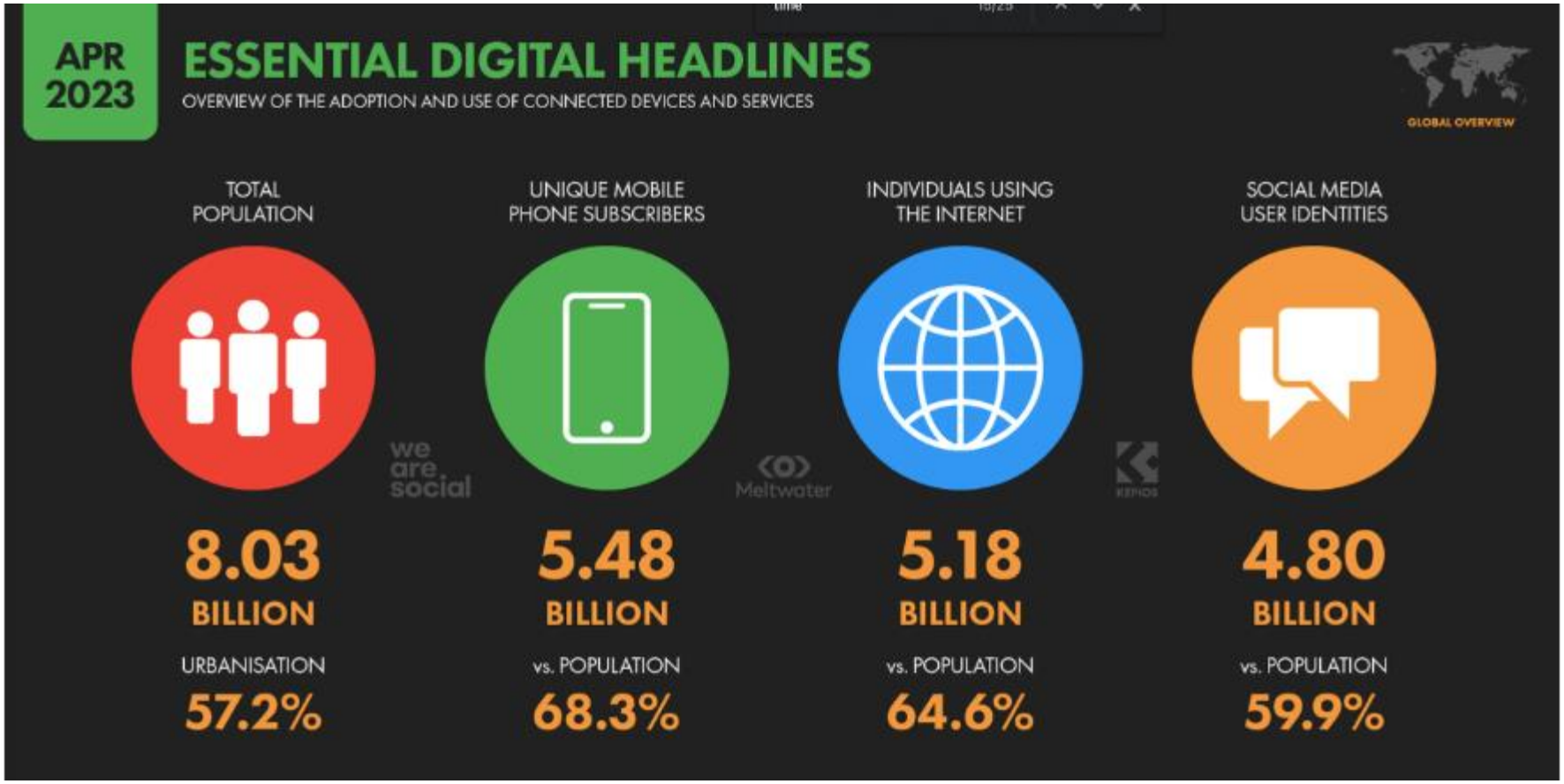
- eBay captures a terabyte of data per minute
- Every mouse click on a web site is captured in Web log files
- Machines (smart meters, Sensors, GPS, etc)
- Social media sites



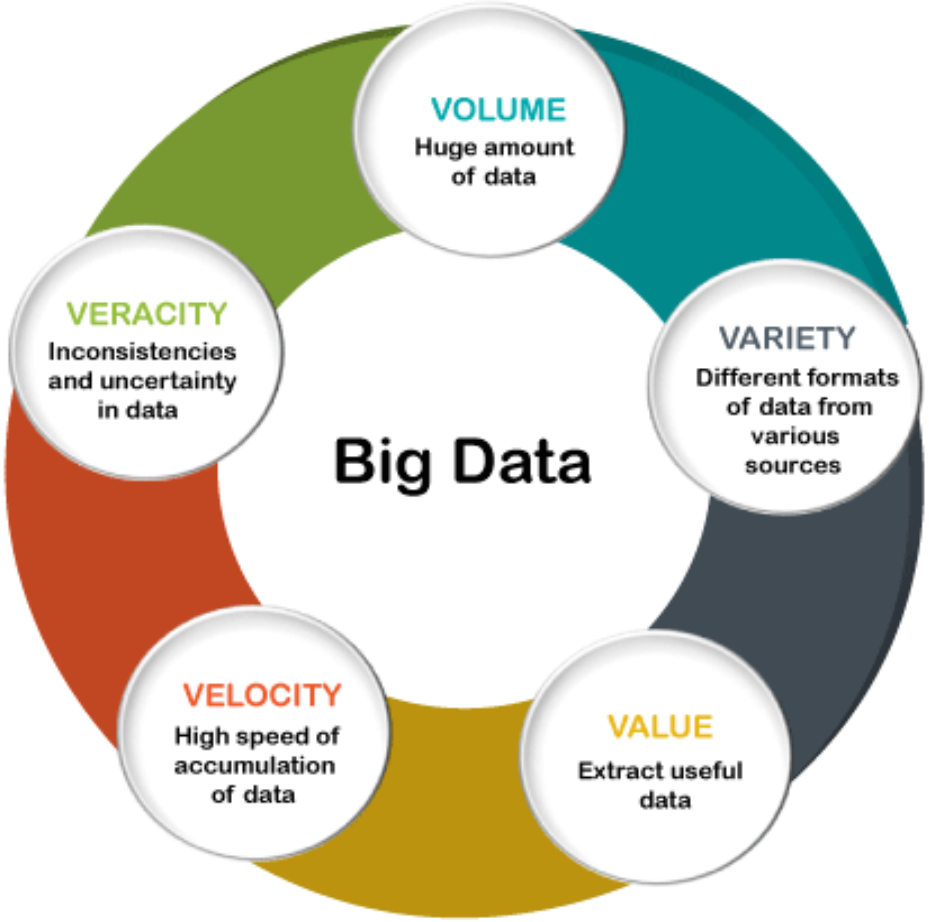
11

# Digital world

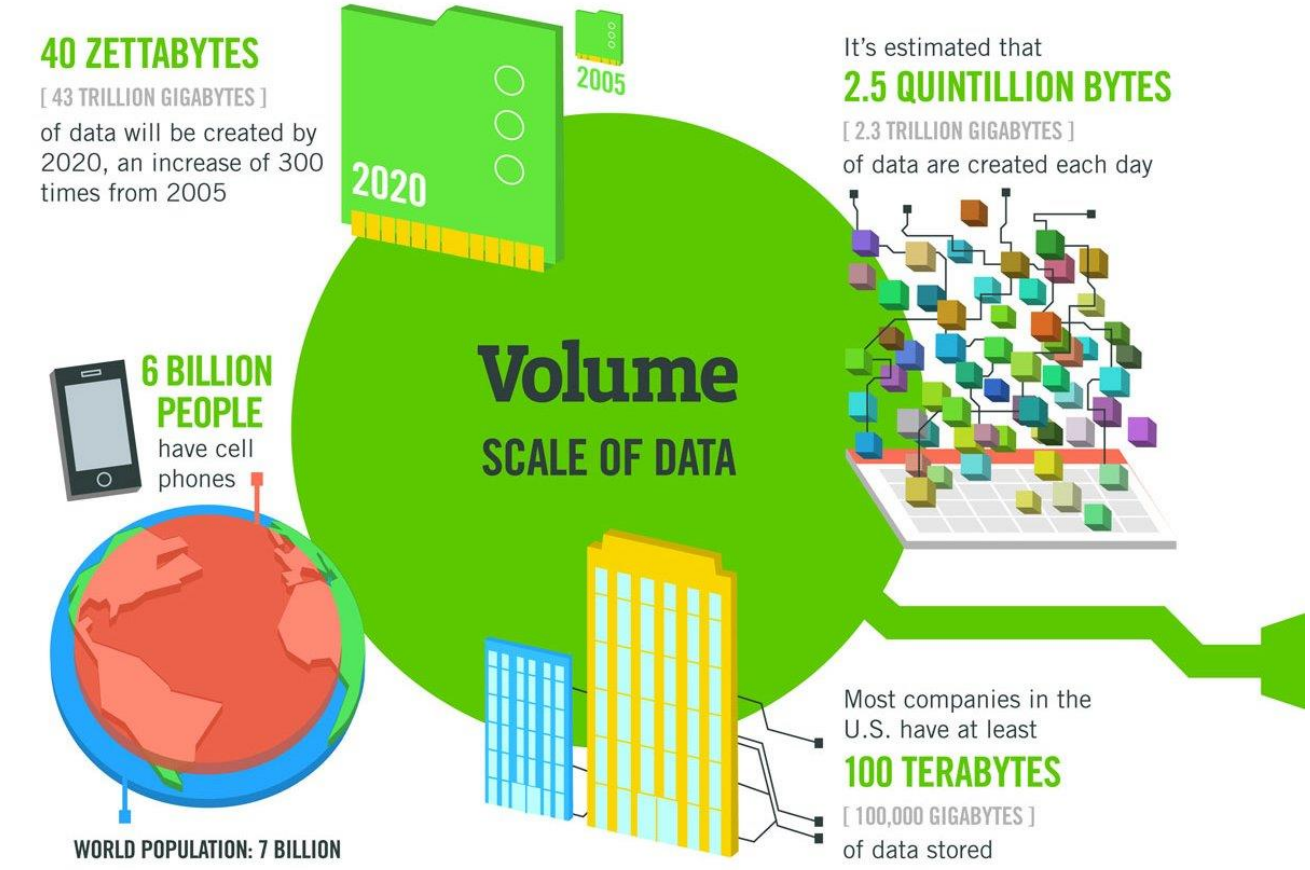


# Characteristics of the Data Revolution

# Characteristics of Big Data



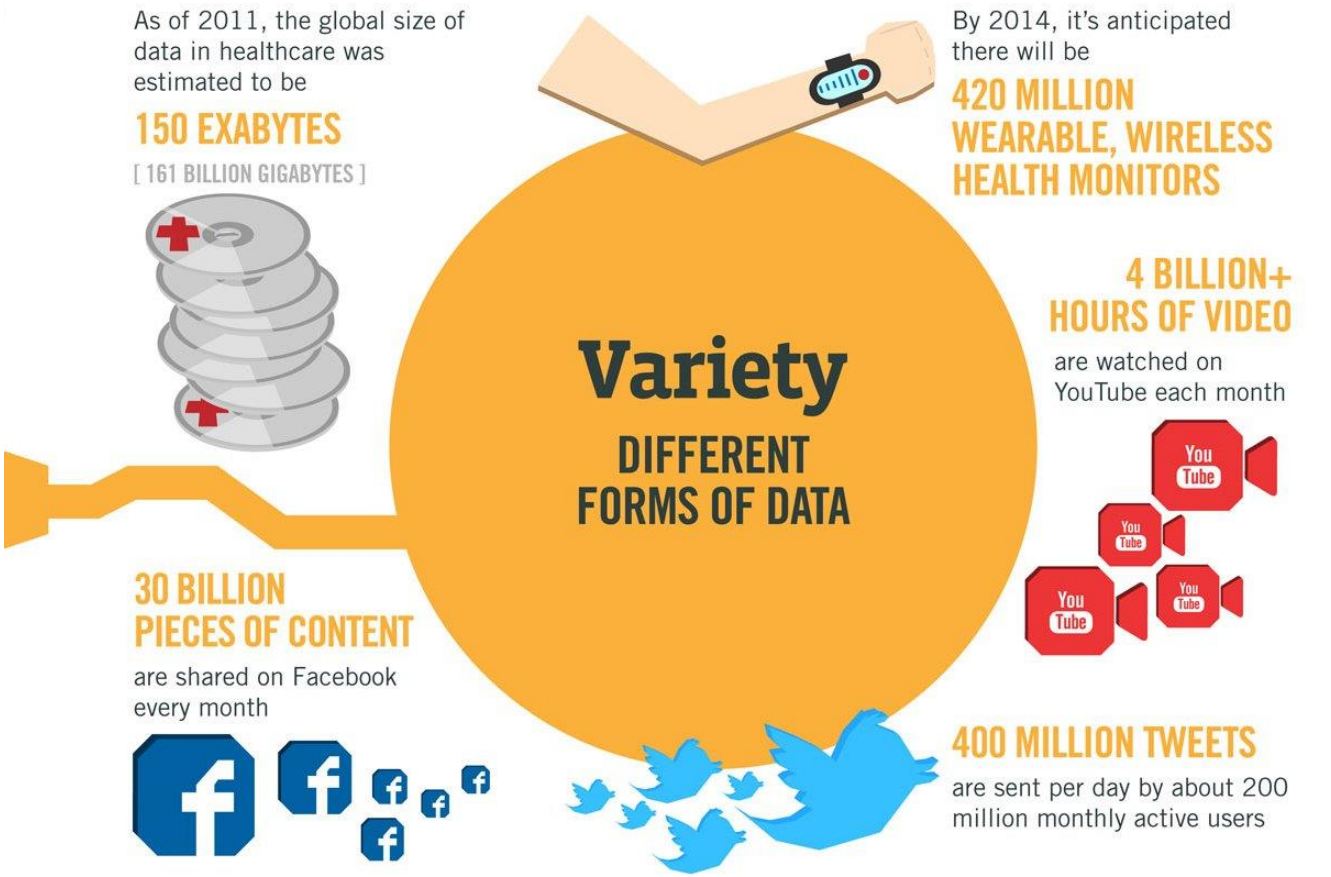
# Volume


40 ZETTABYTES
[ 43 TRILLION GIGABYTES ]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[ 2.3 TRILLION GIGABYTES ]
of data are created each day
6 BILLION PEOPLE
have cell phones
Volume
SCALE OF DATA
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[ 100,000 GIGABYTES ]
of data stored

# Volume

- Many factors contribute to the increase in data volume
  - Transaction-based data stored through the years.
  - Unstructured data streaming in from social media.
  - Increasing amounts of sensor
- Data Volume
  - Growth 40% per year
  - From 8 zettabytes (2016) to >100zb (2023)
  - Data volume is increasing exponentially
  - 90% of the data is created in the past two years

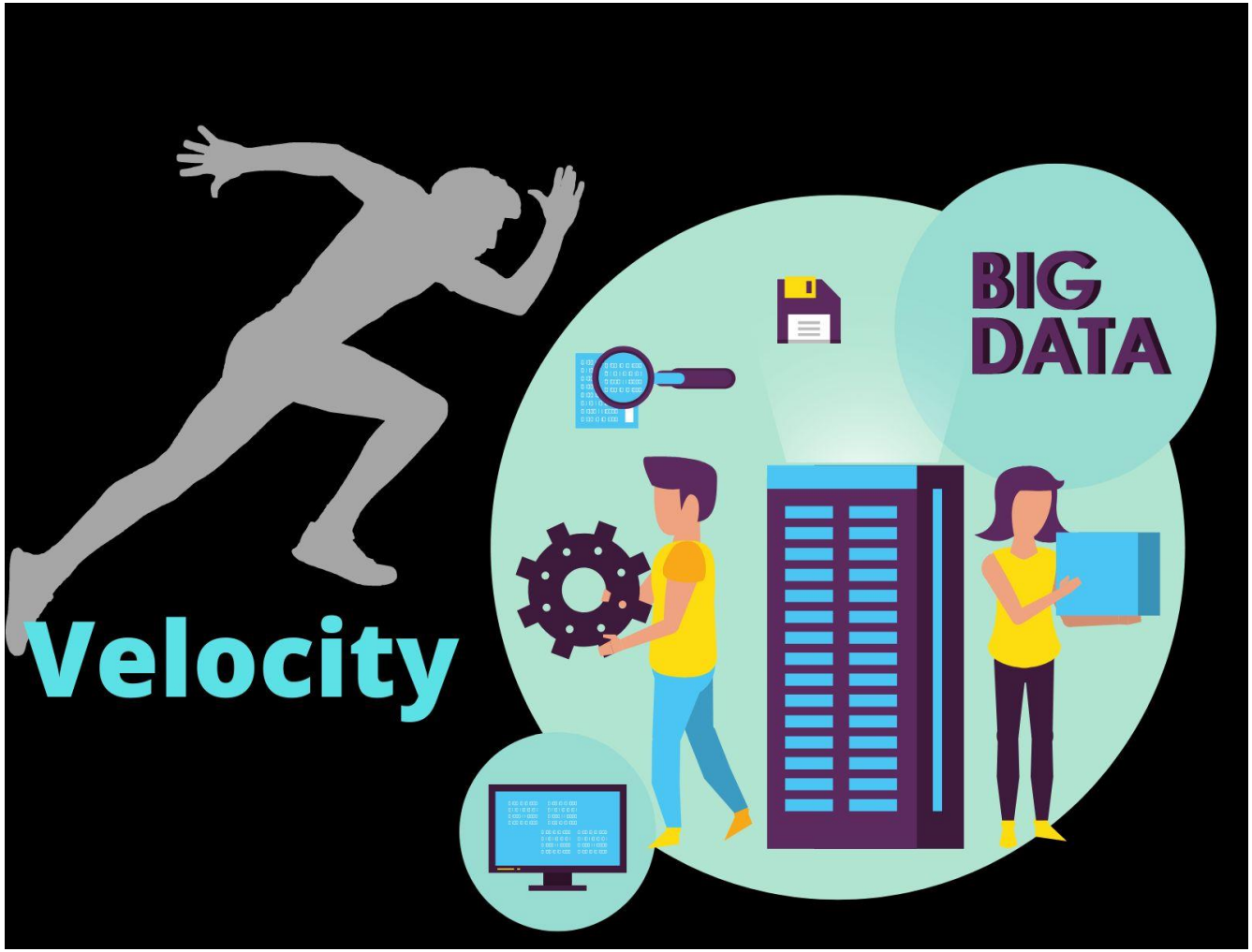
# Variety


As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[ 161 BILLION GIGABYTES ]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

# Variety

- Data today comes in all types of formats.
  - Structured
    - Relational database
  - Semi-structured Data
    - XML
  - Unstructured
    - text documents, email, video, audio
    - Streaming Data
- Different Sources of data is also variety

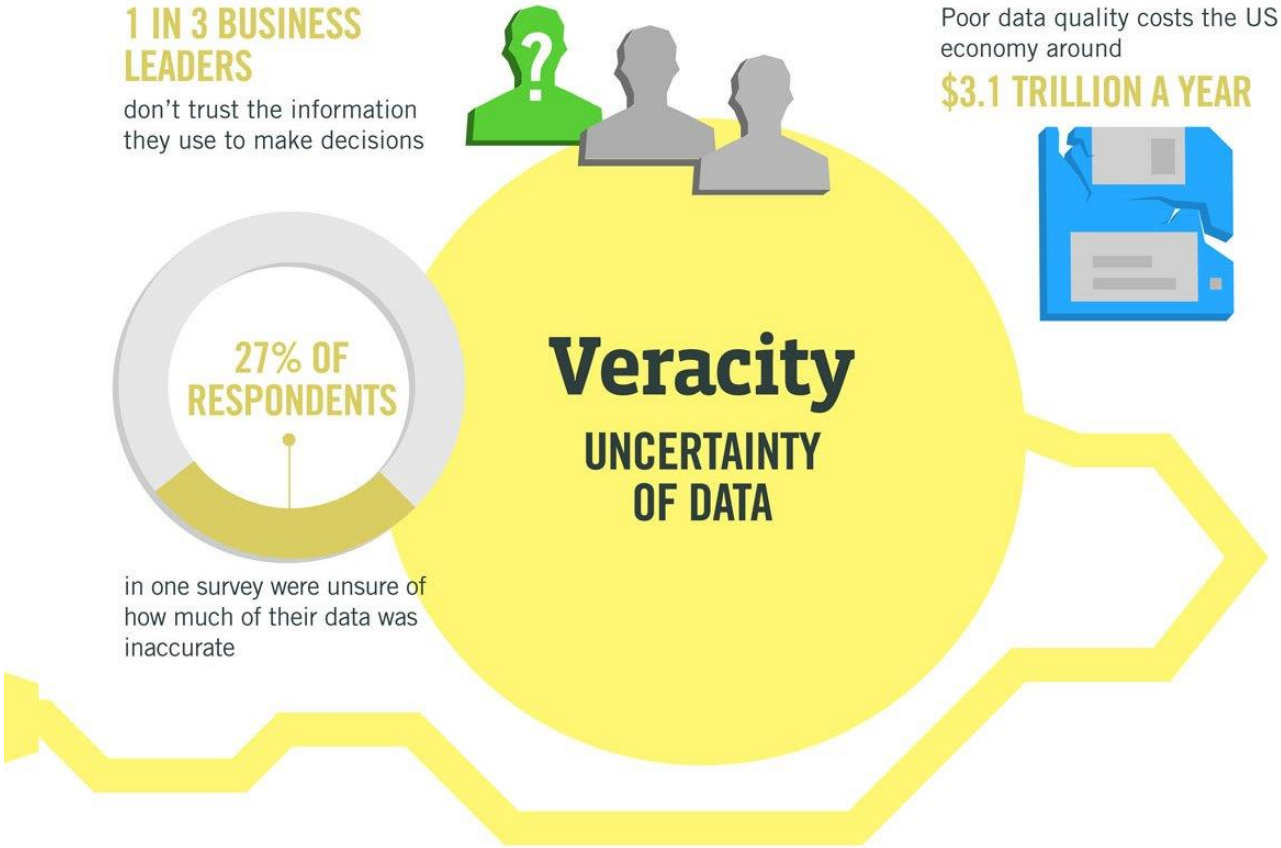
# Velocity



# Velocity

- Data is streaming in at unprecedented speed and must be dealt with timely
- Reacting quickly enough to deal with data velocity is a challenge for most organizations
- Data is being generated fast and need to be processed fast
- Late decisions = missing opportunities
- Examples
  - E-Promotions: Based on the current location, purchase history, what you like, should send promotions timely
  - Healthcare monitoring : sensors monitoring your activities
  - Users comments from social networking sites: must be dealt timely

# Veracity


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

# Veracity

- *Refers to the biases, noise and abnormality in data*
- When we talk about big data, we typically mean its quantity:
  - Is a query feasible on big data within our available resources?
  - How can we make our queries tractable on big data?
- Can we trust the answers to our queries?
  - Dirty data routinely lead to misleading financial reports, strategic business planning decision □ loss of revenue, credibility and customers, disastrous consequences

# Causes of Data Revolution

# Causes of Data Revolution

| Year | Event |
|---|---|
| 1991 | • World Wide Web is born |
| 1995 | • Sun releases the Java platform<br>• Global Positioning System (GPS) |
| 1999 | invents the term - Internet of Things |
| 2001 | Wikipedia is launched |
| 2003 | • The amount of data created surpasses the amount of data created in all of human history before then<br>• LinkedIn launched, 260 million users by 2013 |
| 2004 | Facebook is launched, 1.15 billion user by 2013 |
| 2008 | The number of devices connected to the Internet exceeds the world's population. |
| 2011 | • The IPv4 address space have all been assigned, 4.5 billion unique addresses assigned |
| 2012 | The Obama administration announces the Big Data Research and Development Initiative |

# Causes of Data Revolution

- Major derivers can be identified as the major cause
  - Development of the Web
  - Open data initiatives across the glob
  - Internet of Things

# Web Technologies

- Tim-Berners-Lee invented the Web in 1991
  - The great mind
- We all love him mostly in his AAA Slogan
  - Anyone can say Anything about Any topic
  - His decision in making the web open web
- Tim-Berners-Lee published two articles
  - Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor, 1999
  - The Semantic Web, 2001
- He claim the web that he envision is not realized

# 1. Web Technologies

- Data becomes first class citizen of the Web
- W3C, an authorized body to set standards in the web
  - Data publishing standards
    - HTML,XML,….

# Web Technologies

HTML

```
<p>The Semien Mountains National Park is established in 1969 and made a
UNESCO world heritage site nine years later, the 220km square Semein
Mountains National Park protects the western part of eponymous mountain
range, a serious of incised plateau characterized by sheer 1,000m-high cliffs
and rugged pinnacles and buttresses. </p>
```

XML

```
<?xml version="1.0"? >
<!DOCTYPE "SemienMountains.dtd" >
<SemienMountainPark >
    <FoundedIn>1969 <FoundedIn>
    <RegisteredUNESCO> 1978 <RegisteredUNESCO>
<SemienMountainPark>
```

# Web Technologies: Monetization

- Data becomes an important asset
- Customer experience from one industry is anonymized, packaged, and sold to other industries.
- Internet advertising
  - Yelp lets consumers share their experiences regarding restaurants, shopping, nightlife, beauty spas, active life, coffee and tea, and others.
  - Online platform like amazon ,e-bay capture every activity of users
    - Mouse click, like, post, tweet, query, etc.
    - Every web activity can be sold/bought
  - Google knows more about me than I know

# 2.Open Data initiatives

- The world accepted open data as a tool to fight many problems
- Open government data has got political commitment all over the world
  - A vital communications channel between governments and the public
  - Open data is recognized as a tool to the success of the SDGs
  - G8 leaders signed an Open Data Charter, promising to make public sector data openly available, without charge and in re-useable formats.

# Open Data initiatives

- Promote transparency
- Allow the creation of new, innovative, added-value services
- Accelerate scientific progress
- Improve the quality of decision-making
  - providing the means for evidence-based policy development
- Foster collaboration across government and beyond

# Open Data initiatives cont..

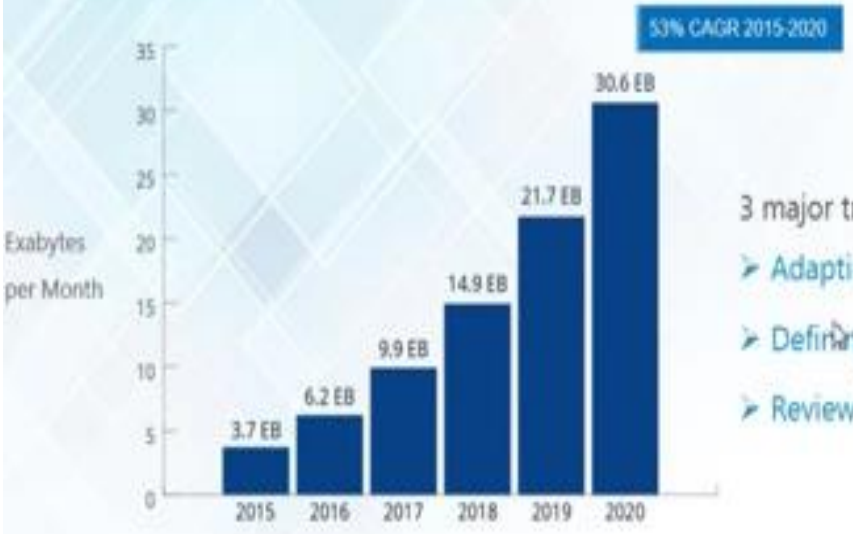
- The open data movement is about making data public so they can be used freely by everyone.
  - US government data
  - UK government data
  - BBC music database
  - General knowledge ontologies such as DBpedia,YAGO and Cyc
  - Various kinds of geographical data e.g., Geonames or OpenStreetMap
  - National library catalogs (USA, Germany etc.)
  - Scientific publications (DBLP)
  - Kenya open data
  - Ethiopian open data under infant stage

# 3. Internet of Things

## Global Mobile Data Traffic, 2015 to 2020

edureka

Cisco Forecasts 30.6 Exabytes per Month of Mobile Data Traffic by 2020




3 major trends contributing to the growth of mobile data traffic:

- Adapting to Smarter Mobile Devices
- Defining Cell Network Advances—2G, 3G, and 4G (5G Perspectives)
- Reviewing Tiered Pricing—Unlimited Data and Shared Plans

Source: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

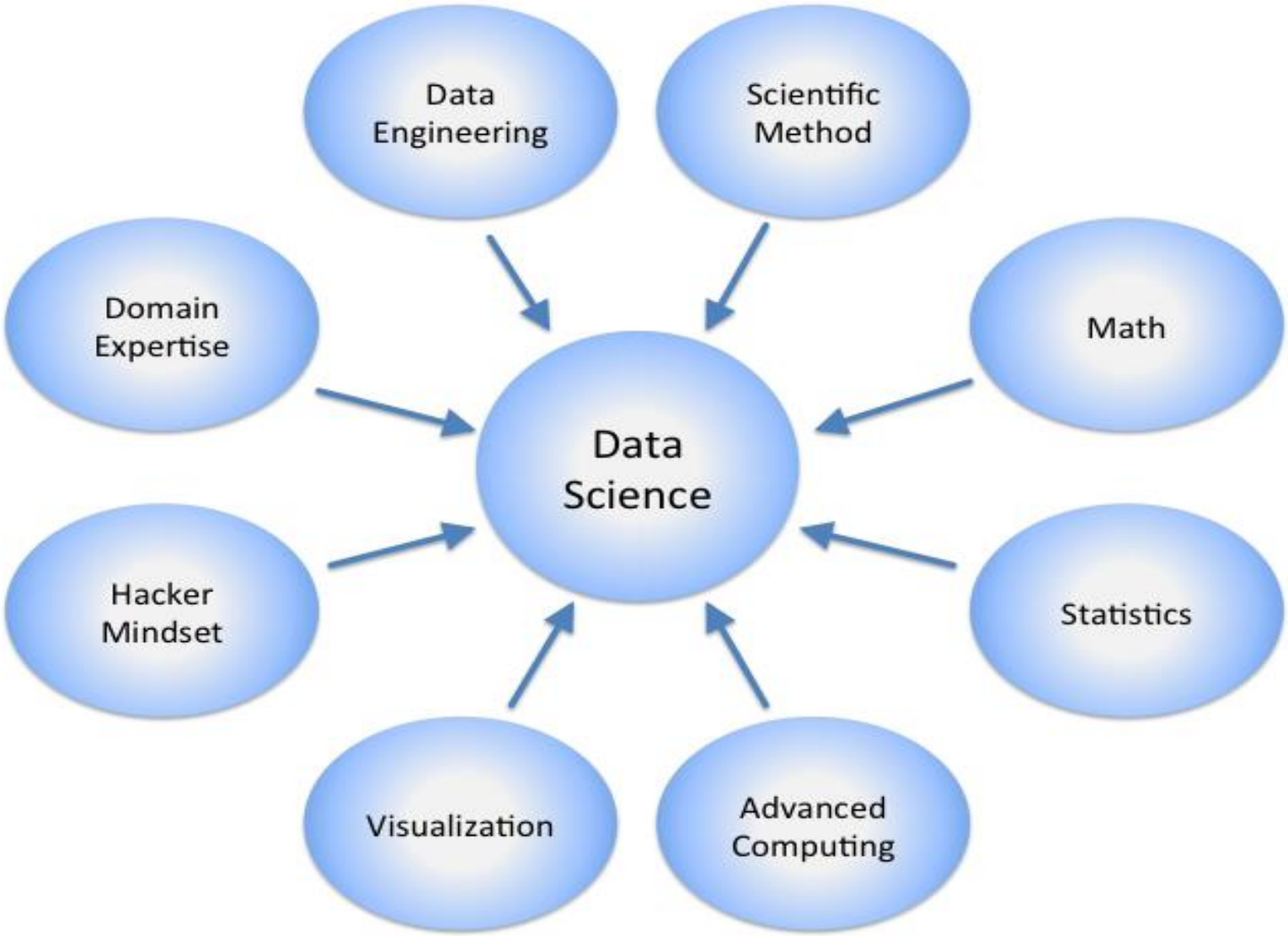
# Definition of Data Science

- Data science incorporates principles, techniques, and methods from many disciplines and domains including data cleansing, data management, analytics, visualization, engineering, and in the context of Big Data.
- Data science combines various technologies, techniques, and theories from various fields, mostly related to computer science and statistics and maths, to obtain actionable knowledge from data.

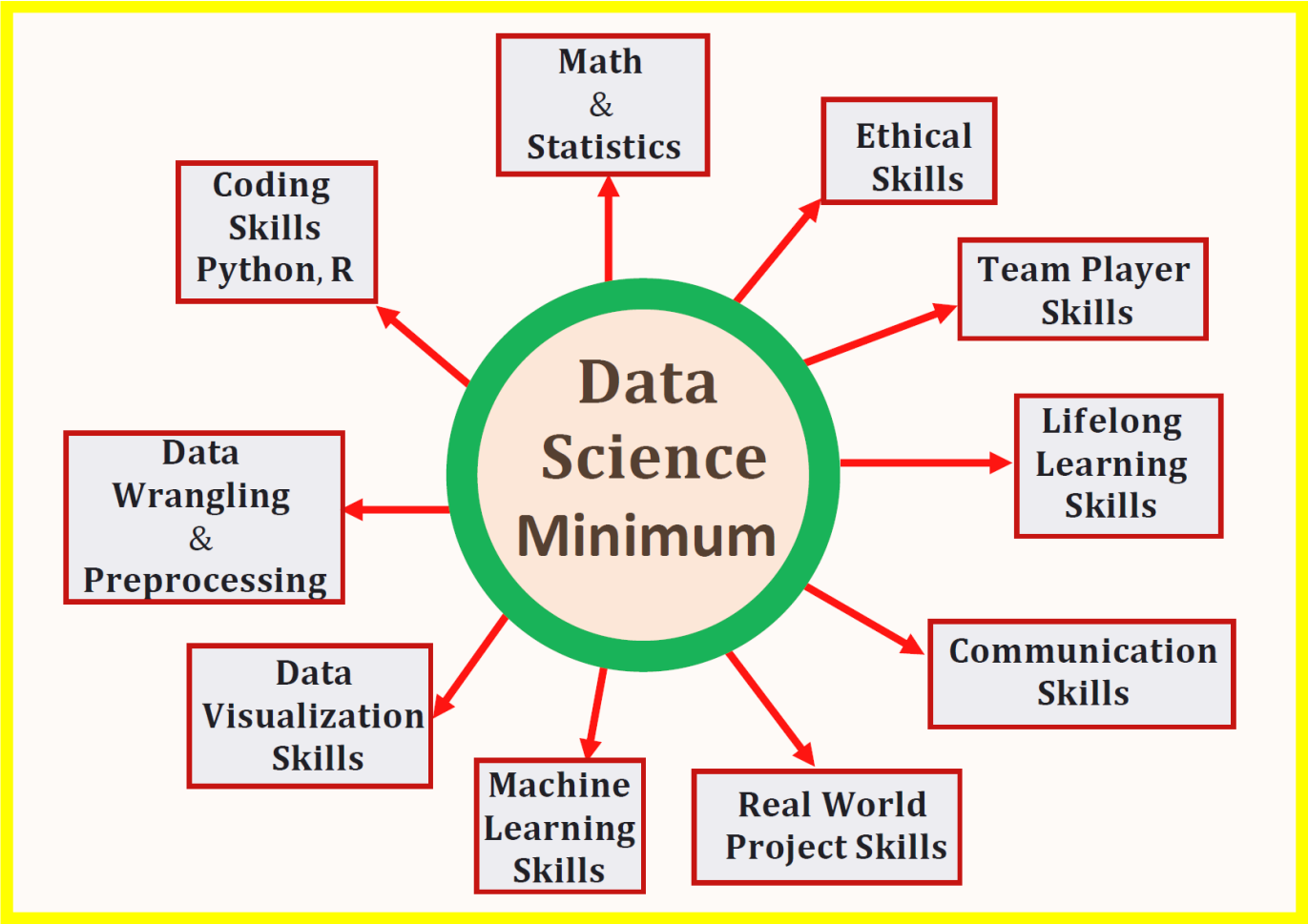
# Definition of Data Science

- **Data Science** is a body of principles and techniques for applying data-intensive analysis to investigate new knowledge from data .
- In simple terms, it is the umbrella of techniques used when trying to extract insights and information from data.
  - To see patterns
  - To discover relationships
  - To make sense of data

# Data Science



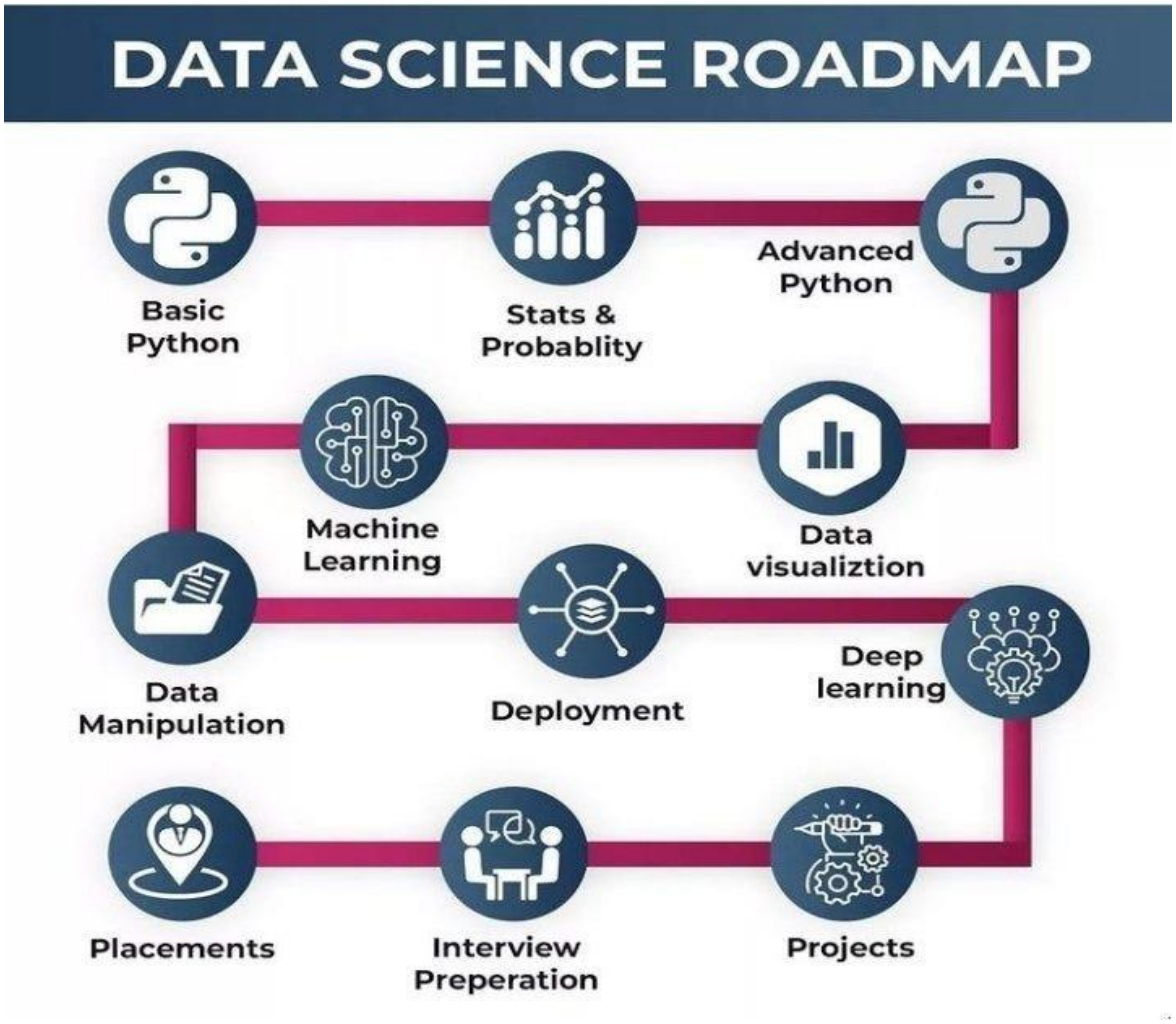
# Identified Data Science *Skills*



# Data scientist path

- Learn Data Science Fundamentals
- Learn Key Programming Languages for Data Science
- Learn how to do visualizations
- Work on some Data Science projects that will help develop your practical data skills
- Make a Portfolio that shows your Data Science Skills
- Go for job

# Data science road map



# Demand for Data Science

- According to US News and World Report in 2023, information security analyst, software developer, data scientist ranked among the top jobs in terms of pay and demand

**Data scientist**

- **Average annual salary:** $152,279

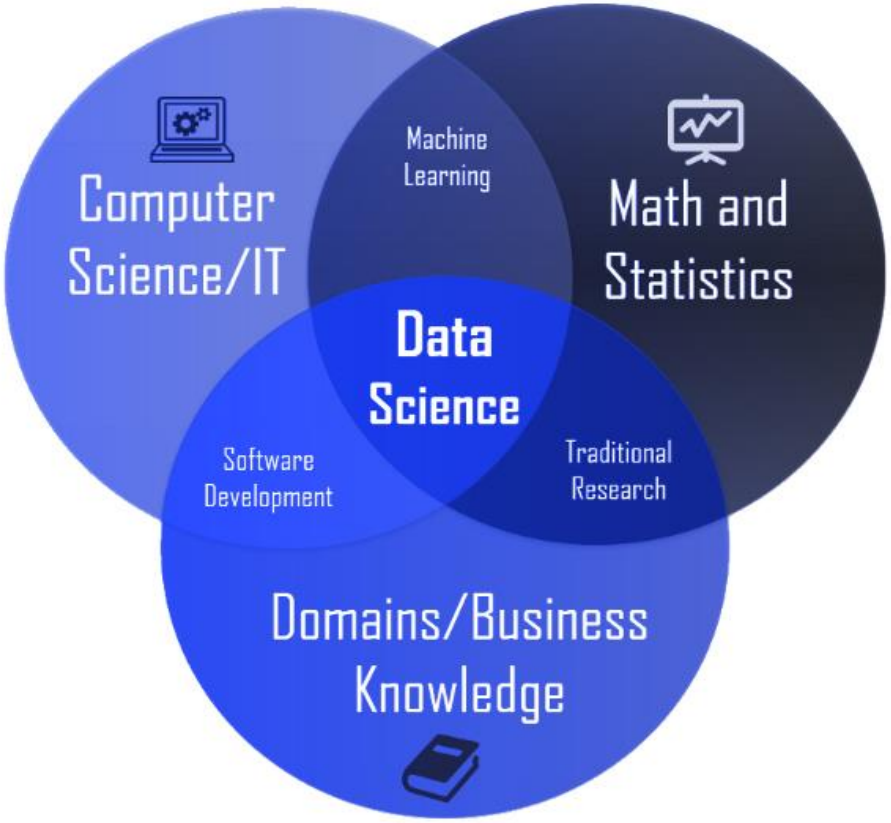
# Chapter -two

# DS and its relation with other disciplines , application areas

# Data Science and its relation to others

- The field of Data Science is quite a huge one and it has various branches.
- ranges from when data is being collected to analyzes and presentation (visualization),prediction of results.
- The Data Science process involves different skill-sets and disciplines for efficiency and effectiveness

# Data Science and its relation to others



# DS vs STATISTICS

•In general, statistics is the study of numerical or quantitative data to make predictions or draw conclusions about a population.

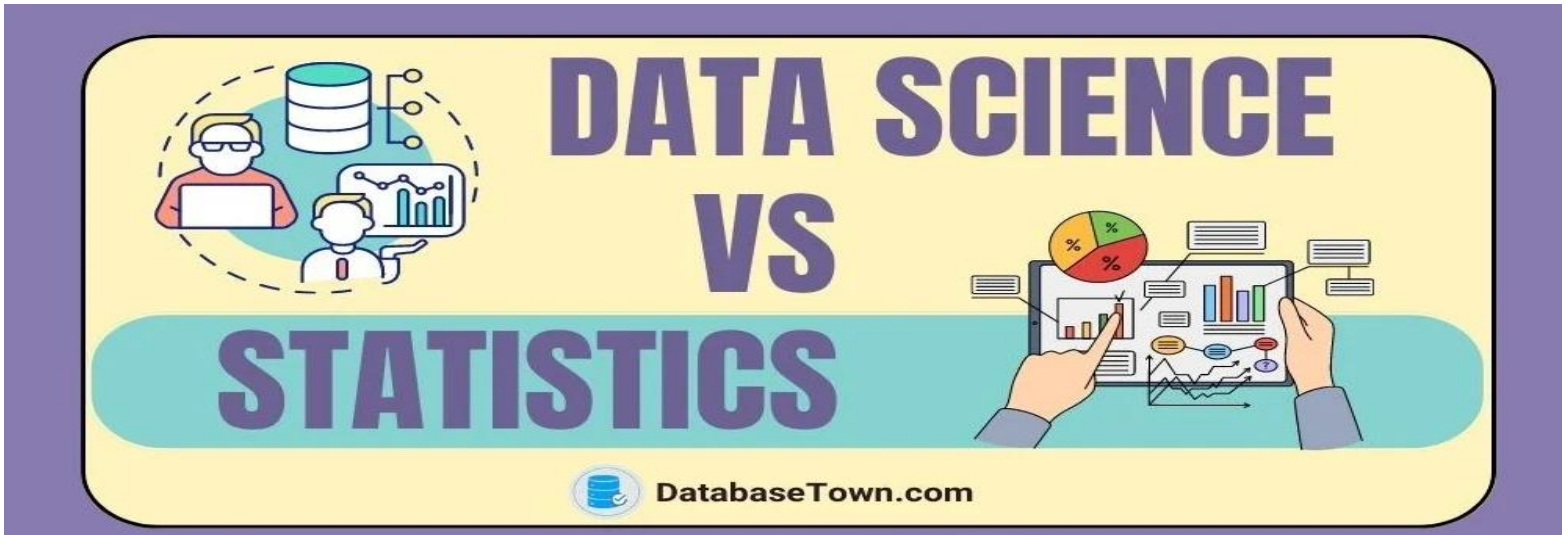
- Data science uses statistical methods, Mathematical and programming skills to analyze large amounts of data and understand the results better

•Statistics education deals traditionally with (small enterprises):

**Structured data**

- •Data from sampling/sensus
- •Inferential study (estimation and hypotheses testing)

# DS vs STATISTICS




| Criteria | Data Science | Statistics |
|---|---|---|
| Definition | Data science is field where data scientists extract useful information from raw data using different methods. | Statistics is the mathematical discipline of collecting, analyzing, and interpreting data to make predictions. |
| Scope and Objectives | Broad scope, it covers data collection, analysis, and modeling. | Primarily focused on data analysis, inference, and probability. |

# Data Management Systems

- Relational Database Management systems
  - Good at structured data, form of table
  - They have limitation on
    - unstructured, quasi- or semi-structured data
  - Many insights could be extracted from the unstructured, quasi- or semi-structured data

| | Databases | Data Science |
|---|---|---|
| Data Value | Precious | Cheap |
| Data Volume | Modest | Massive |
| Examples | Bank records,<br>Personnel records,<br>Census,<br>Medical records | Online clicks,<br>GPS logs,<br>Tweets,<br>Big data,<br>sensor readings |
| Priorities | Consistency,<br>Error recovery,<br>Auditability | Speed,<br>Availability,<br>Query richness |
| Structured | Strongly (Schema) | Weakly or none (Text) |

| Databases | Data Science |
|---|---|
| Querying the past | Querying the future |

# Machine Learning

**Data Science systems**

- Collection and processing of data
- Data visualization – Visually explore data to get a better intuition of data
- Data engineering – Making sure hot and cold data is always accessible. Covers data backup, security, disaster recovery
- Deployment in production mode – Migrate system into production with industry standard practices.
- Automated decisions – This includes running business logic on top of data or a complex mathematical model trained using any ML algorithm.

**Machine Learning modeling**

- Understand problem
- Explore Data
- Prepare data
- Select a model and train
- Deploye

# Data Science Discipline Knowledge Areas

# Data Scientists Data Engineer and Analyst

# Societal Problems Addressed

# HealthCare

- **Medical Image Analysis**
  - Procedures such as detecting tumors, artery stenosis, lung texture classification
- **Genetics & Genomics**
  - Understand the impact of the DNA on our health and find individual biological connections between genetics, diseases, and drug response

# Retail

- Customer is savvy, impatient and busy.
- They want instant gratification and excellent customer service.
- In order to compete and stay one step ahead, retailers need to have a 360-degree view of the customer.
- Retail analytics helps businesses get deep insights into customer behavior and act accordingly
  - Identify items that are likely to be purchased together.
  - Which marketing strategies work better than others?
  - Optimal Pricing
  - What promotions and offers to employ in each store?
  - Store wise product-mix
  - Personalized offers
  - Efficient stock strategy

# E-commerce

- E- commerce businesses primarily use analytics to understand:
  - Acquisition - how your visitors and customers found and arrived at your site.
  - Shopping and purchasing behavior: how users engage with your website, which products they view, which ones they add or remove from shopping carts; along with initiating, abandoning, and completing transactions.
  - Economic Performance – how many products the average transaction includes, the average order value, refunds you had to issue.

# Finance

- The global financial analytics market is one of the fastest growing sectors of the data industry.
- Organizations big and small are investing in financial analytics tools and technologies to solve specific business problems, reduce costs, improve budgets and get insights into future financial scenarios.
- Typically financial analytics includes
  - Risk analysis
  - Working capital management
  - Fraud detection and prevention

# Healthcare, Education, Telecom etc

- Analytics can be used for evidence based medical care, improved patient care, predicting outbreaks of diseases and reducing hospital operating costs.

- Analytics is also being used to improve teaching practices. It also enables teachers to better monitor student progress, personalize learning and improve educational institutions operational efficiencies.

- In the telecom industry analytics is fast gaining much ground. Operators are using analytics to drive revenue,and improve network performance.

# Marketing

- Understanding customers and how to find more people like them is the key to sustainable growth.
- It helps measure, manage and analyze marketing performance to maximize its effectiveness and optimize return on investment (ROI).
  - How are our marketing initiatives performing today?
  - How can we improve those which are not effective?
  - How do our marketing activities compare with our competitors'?
  - What can we learn from our competition?
  - Are our marketing resources properly allocated?
  - Are we using the right channels?